

کاهش شکاف معنایی در دسته‌بندی پرسش‌ها با بهره‌گیری از قوانین طبقه‌بندی

سیده زهرا آفتابی^۱، دانشجو، محمدعلی زارع چاهوکی^۲، استادیار

۱- دانشکده مهندسی برق و کامپیوتر - دانشگاه یزد - یزد - ایران - szaftabi@stu.yazd.ac.ir

۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه یزد - یزد - ایران - chahooki@yazd.ac.ir

چکیده: دسته‌بندی پرسش‌ها^۱ یکی از مؤلفه‌های حیاتی سیستم‌های بازیابی اطلاعات^۲ و پاسخ‌گویی به پرسش^۳ است. هدف از دسته‌بندی پرسش، شناسایی دقیق نوع پاسخ موردانتظار آن و انتساب برجستگی به آن مطابق با دسته‌ای است که پرسش در آن قرار می‌گیرد. تاکنون با دو رویکرد مبتنی بر قانون^۴ و یادگیری ماشین^۵، پژوهش‌های متعددی در این حوزه صورت پذیرفته است. هدف ما در این پژوهش تلفیق نتایج این دو رویکرد به منظور افزایش صحت^۶ دسته‌بندی است. نوآوری اصلی ارائه‌شده در این پژوهش، غنی‌سازی بردار ویژگی کلمات^۷ حاصل از پرسش‌ها با قوانین دسته‌بندی است. اهمیت روش تلفیق ارائه‌شده در این مقاله امکان استفاده از مخازن قوانین با ساختار طبقه‌بندی متفاوت نسبت به ساختار موجود برای دسته‌بندی پرسش‌ها است. نتایج حاصل از پیاده‌سازی روش پیشنهادی بر دادگان UIUC بیانگر مؤثر بودن روش پیشنهادی در بهبود صحت دسته‌بندی پرسش‌ها است.

واژه‌های کلیدی: دسته‌بندی پرسش، رویکرد مبتنی بر قانون، رویکرد مبتنی بر یادگیری ماشین، رویکرد ترکیبی.

Bridging the semantic gap in question classification by categorization rules

S. Z. Aftabi, Student¹, M. A. Zare Chahooki, Assistant Professor²

1- Faculty of Electrical and Computer Engineering, University of Yazd, Yazd, Iran, Email: szaftabi@stu.yazd.ac.ir

2- Faculty of Electrical and Computer Engineering, University of Yazd, Yazd, Iran, Email: chahooki@yazd.ac.ir

Abstract: Question classification is one of the vital components of information retrieval and question answering systems. The goal of question classification is to accurately identify the expected answer type of question and assign a label to it based on the class that the question placed in. So far, several studies have been done in this area using rule-based methods and machine learning techniques. The objective of this study is to combine results of these two approaches in order to enhance the accuracy of classification. The main innovation in this research is the enrichment of the bag of words feature vector of categorization rules. The importance of the integrated method presented in this paper is the use of rule repository with different categorization structure than the existing structure for question classification. Experimental results of the proposed method on UIUC dataset show its effectiveness on improving the question classification accuracy.

Keywords: Question classification, rule-based method, machine learning method, hybrid method

تاریخ ارسال مقاله: ۱۳۹۴/۹/۱۶

تاریخ اصلاح مقاله: ۱۳۹۴/۵/۲۱ و ۱۳۹۴/۶/۲۱

تاریخ پذیرش مقاله: ۱۳۹۴/۹/۱۷

نام نویسنده مسئول: محمدعلی زارع چاهوکی

نشانی نویسنده مسئول: ایران - تبریز - یلوار ۲۹ بهمن - دانشگاه تبریز - دانشکده مهندسی برق و کامپیوتر

۱- مقدمه

انواع پرسش نباشد. علاوه بر این، یک مجموعه از قوانین، ممکن است روی یک دادگان، نتایج بسیار خوبی ارائه دهد ولی برای دادگانی دیگر عملکرد ضعیفی داشته باشد. ضمن اینکه با تغییر دسته‌بندی، مجموعه قبلی دیگر کارآمد نخواهد بود. همچنین اگرچه اغلب کاربران پرسش‌های خود را در قالب‌های رایج What، When، Where، Why، How و Which ارسال می‌کنند، اما برخی نیز از این الگوها پیروی نمی‌کنند. بنابراین هر چند برخی از پرسش‌های کاربران مضمون یکسان دارند اما به گونه‌های متفاوتی بیان می‌شوند. لذا گردآوری مجموعه‌ای از قوانین به طوری که تمام صورت‌های مختلف پرسش را دربرگیرد امری دشوار و زمان‌بر است [۵، ۷]. در این راستا تاکنون پژوهش‌های [۶، ۸، ۹، ۱۵] از مجموعه‌ای پیش‌ساخته از قوانین استفاده نموده‌اند.

رویکرد مبتنی بر یادگیری ابتدا به تجزیه و تحلیل پرسش‌های برچسب‌دار پرداخته و الگوهای را بر اساس آن می‌آموزند. به این ترتیب، نوع پاسخ هر پرسش بدون برچسب ورودی را مطابق با آموخته‌های خود پیش‌بینی می‌کند [۵، ۷، ۱۴]. با توجه به اینکه روش‌های مبتنی بر یادگیری ماشین دارای قابلیت شناسایی ساختارهای نهفته موجود در نمونه‌های برچسب‌دار هستند، اکثر پژوهش‌های اخیر در حوزه دسته‌بندی پرسش، استفاده از این روش‌ها را در اولویت قرار داده‌اند. این پژوهش‌ها در مدل دسته‌بندی یا روش استخراج ویژگی از نمونه‌ها با یکدیگر متفاوت هستند [۵، ۷، ۱۴]. انتخاب مدل یادگیری ماشین، صحت فرآیند دسته‌بندی را تحت تأثیر قرار می‌دهد [۲]. در این راستا، تاکنون از مدل‌های ماشین بردار پشتیبان^۸ (SVM) [۴، ۶، ۹-۱۲]، درخت تصمیم^۹ (DT) [۱۱]، بیشینه بی‌نظمی (ME) [۹، ۱۱]، بیزین^{۱۰} (NB) [۴، ۱۱]، شبکه‌های عصبی (NN) [۱۶، ۱۷] و شبکه پراکنده غربالگری^{۱۱} (SNoW) [۱۳] استفاده شده است.

در حوزه دسته‌بندی پرسش، ترکیب روش‌های مبتنی بر یادگیری و قانون با یکدیگر دارای اهمیت است. زیرا می‌توان از دستاوردهای هر دو رویکرد در بهبود صحت فرآیند دسته‌بندی بهره برد. هانگ و همکارانش در [۹] برای پیاده‌سازی دسته‌بندی خود از دو الگوریتم SVM و ME استفاده نمودند. آن‌ها ترکیبات مختلف تک-گرام، دو-گرام و سه-گرام‌های هر پرسش را همراه با کلمه پرسشی، سرواژه^{۱۲} پرسش و کلماتی را که سرواژه نوع خاصی از آن‌ها محسوب می‌شود^{۱۳}، به عنوان ویژگی‌های بردار ویژگی نمونه‌ها استخراج نمودند. به منظور شناسایی سرواژه هر نمونه، مجموعه‌ای از قوانین را در قالب عبارتهای مستقل از متن به کار گرفتند تا بتوانند بخش‌هایی از جمله را که احتمال دارد سرواژه در آن‌ها قرار گرفته باشد، شناسایی نمایند. این مجموعه با عنوان قوانین کالینز نام‌گذاری شده است. البته قوانین کالینز در بسیاری از اوقات نمی‌توانند سرواژه را به درستی تشخیص دهند؛ زیرا آن‌ها عبارتهای فعلی را نسبت به عبارتهای اسمی در اولویت قرار می‌دهند. در حالی که سرواژه اکثر جملات پرسشی در عبارتهای اسمی

امروزه وب، حجم عظیمی از اطلاعات و دانش را در خود گنجانده است. با افزایش روبه‌رشد مخازن اطلاعات و دانش در وب، نیاز به ساخت موتورهای جستجوی هوشمند که قادر به پاسخ‌گویی به پرسش‌های کاربران نیز باشند کاملاً احساس می‌شود. به این ترتیب، سیستم‌های پاسخ‌گویی به پرسش (QA) طراحی و پیاده‌سازی شده‌اند. این سیستم‌ها موظف هستند که برای پرسش‌های کاربر که به زبان طبیعی ارسال شده‌اند، پاسخی کوتاه و کامل ارائه دهند [۴-۱]. ساختار پایه یک سیستم پاسخ‌گویی به پرسش از سه مؤلفه "پردازش پرسش"، "بازیابی قطعه‌متون" و "پردازش پاسخ" تشکیل شده است که هر کدام دارای مؤلفه‌ای اصلی و تعدادی مؤلفه جانبی هستند [۲]. اصلی‌ترین مؤلفه در پردازش پرسش، دسته‌بندی آن است. این مؤلفه موظف است نوع پاسخ مورد انتظار پرسش را شناسایی نماید [۵]. به بیان دیگر، هر پرسش را بر اساس نوع پاسخ آن در یکی از دسته‌های از پیش تعریف‌شده قرار می‌دهد.

صحت عملکرد مؤلفه دسته‌بندی پرسش، تعیین‌کننده کیفیت و کارایی سیستم پاسخ‌گویی به پرسش است. وجود این مؤلفه به سه دلیل حائز اهمیت است [۵، ۶]:

الف) پس از شناسایی نوع پاسخ مورد انتظار پرسش می‌توان پرسش را به مناسب‌ترین صورت جهت ارسال به موتور جستجو، اصلاح نمود. همچنین می‌توان مطابق با نوع پاسخ، استراتژی مناسب را برای جستجو برگزید.

ب) با شناسایی نوع پاسخ پرسش می‌توان بافت متنی را که پاسخ در آن ظاهر می‌شود تعیین نمود [۷]. بنابراین فضای جستجوی مورد نیاز برای یافتن پاسخ‌های کاندید بسیار محدود می‌شود. لذا دسته‌بندی نادرست پرسش می‌تواند توانایی رسیدن به یک پاسخ صحیح را سلب نماید.

ج) می‌توان پاسخ‌های کاندید یافت‌شده را بر اساس میزان شباهت آن‌ها به دسته‌ای که پرسش در آن قرار گرفته رتبه‌بندی کرده و پاسخ‌های با بیش‌ترین امتیاز را انتخاب نمود.

برای دسته‌بندی پرسش‌ها، دو رویکرد متفاوت (۱) مبتنی بر قانون [۳، ۶، ۸] و (۲) مبتنی بر یادگیری ماشین [۴، ۶، ۹-۱۳] وجود دارد. اکثر سیستم‌های اولیه پاسخ‌گویی به پرسش، در مؤلفه دسته‌بندی پرسش خود، از مجموعه‌های پیش‌ساخته از قوانین که توسط گروه‌های مختلفی از پژوهشگران گردآوری شده بودند، استفاده می‌کردند [۵، ۶، ۱۴]. برای ساخت این مجموعه از قوانین از انواع مختلفی از کلمات پرسشی، ترکیب‌های کلمات و ویژگی‌های مختلف جمله استفاده شده بود تا ساختار پرسش‌های مختلف موجود در یک دسته معنایی را با عبارتهای منظم مدل نمایند [۱۴].

استفاده از مجموعه قوانین، اگرچه در ظاهر، بسیار ساده به نظر می‌رسد اما با چالش‌های فراوانی روبرو است. چنانچه مجموعه قوانین مورد نیاز به اندازه کافی بزرگ نباشد، ممکن است سیستم قادر به پوشش

رویکرد پیشنهادی این مقاله را به تفصیل معرفی خواهیم نمود. در بخش ۴ اقدامات انجام‌شده به‌منظور ارزیابی و آزمایش رویکرد پیشنهادی تشریح خواهد شد و اشاره‌ای به نتایج سایر پژوهش‌ها خواهیم داشت. در آخر نیز در بخش ۵ به ذکر پیشنهادها برای پژوهش‌های آینده خواهیم پرداخت.

۲- مروری بر دیگر پژوهش‌ها

مؤلفه دسته‌بندی پرسش مهم‌ترین مؤلفه سیستم‌های پاسخ‌گویی به پرسش است. از این‌رو، بسیاری از پژوهشگران در تلاش هستند تا پاسخ مناسبی برای چالش‌های موجود در این حوزه بیابند. از جمله چالش‌های این حوزه که در این پژوهش نیز به آن‌ها پرداخته شده است می‌توان به موارد زیر اشاره کرد [۵]:

(الف) از یک پرسش که به زبان طبیعی بیان شده است چه ویژگی‌هایی را می‌توان استخراج نمود؟

(ب) نقش هریک از ویژگی‌ها در صحت عملکرد دسته‌بند چقدر است؟

(ج) استخراج کدامیک از ویژگی‌ها با توجه به هزینه زمان و امکانات جانبی موردنیاز، به‌صرفه خواهد بود؟

(د) چگونه می‌توان ویژگی‌ها را به طریقی مؤثر با یکدیگر ترکیب کرد؟

(ه) آیا می‌توان از ابزار دیگری مانند وردنت برای درک بهتر معنای واقعی پرسش استفاده نمود؟

(و) آیا می‌توان با بهره‌گیری از مجموعه‌ای از قوانین ساختاری، درک بهتری از معنای پرسش داشت؟

(ز) چگونه می‌توان با استفاده از مجموعه‌ای از پرسش‌های برچسب‌دار، مدلی را آموزش داد که بتواند پرسش‌های بدون برچسب را با صحت بالا دسته‌بندی نماید؟

(ح) استفاده از کدام دسته‌بند مبتنی بر یادگیری، مناسب‌تر خواهد بود؟
(ط) چگونه می‌توان از دستاوردهای دورویکرد مبتنی بر قانون و یادگیری ماشین به‌عنوان مکمل یکدیگر بهره برد؟

محققان تاکنون دو رویکرد مبتنی بر قانون و مبتنی بر یادگیری ماشین را برای دسته‌بندی پرسش‌ها و یافتن پاسخ مناسب به چالش‌های فوق به کار گرفته‌اند.

۲-۱ - روش‌های مبتنی بر قانون

اکثر سیستم‌های اولیه پاسخ‌گویی به پرسش، در مؤلفه دسته‌بندی پرسش خود، از مجموعه‌های پیش‌ساخته از قوانین استفاده کرده‌اند. مجموعه قوانینی که آن‌ها مورد استفاده قرار داده‌اند از عبارت‌های منظم به‌منظور مدل‌سازی ساختار پرسش‌های مختلف موجود در دسته‌های معنایی استفاده نموده‌اند. از آن جمله می‌توان به پژوهش [۸] اشاره نمود. هرم جیکوب در این پژوهش، ۱۷۳۸۴ پرسش همراه با پاسخ‌های آن‌ها که از سایت answer.com دریافت شده بودند تجزیه و تحلیل نموده است. سپس بر اساس ساختار پرسش‌هایی که در یک دسته معنایی به لحاظ نوع پاسخ مورد انتظارشان قرار گرفته‌اند، مجموعه‌ای از قوانین ساختاری گردآوری کرده است. این مجموعه

جمله‌گنجانده شده است. از این‌رو، آن‌ها در پژوهش خود، تعدادی از این قوانین را تغییر دادند [۷].

سیلوا و همکارانش در [۶] از رویکرد ترکیبی استفاده نمودند. آن‌ها ابتدا در فاز اول، برچسب‌دهی به پرسش بر اساس انجام تطابق مستقیم بین پرسش و مجموعه‌ای از ۶۰ قانون را انجام دادند. چنانچه برای یک پرسش هیچ تطابقی یافت نمی‌شد ابتدا از طریق تطابق درخت تجزیه پرسش با نسخه تغییر یافته قوانین کالینز، سرواژه پرسش را شناسایی می‌کردند. سپس با استفاده از وردنت^{۱۴}، در میان کلمات با معنای عام‌تر عام‌تر که سرواژه، نوع خاصی از آن‌ها محسوب می‌شود، تا عمق مشخصی پیش رفته و نام دسته‌ای را که بیش‌ترین شباهت معنایی را به سرواژه پرسش داشت برمی‌گزیدند. ایشان از خروجی فاز اول به‌عنوان ویژگی جدید در بردار ویژگی پرسش جهت ارسال به الگوریتم یادگیری ماشین استفاده نمودند. نتایج این پژوهش نشان داد که ترکیب دسته‌بندی مبتنی بر قانون با یادگیری ماشین نسبت به دسته‌بندی مبتنی بر قانون به‌تنهایی، صحت فرآیند دسته‌بندی را بهبود می‌دهد [۵، ۷].

در این مقاله نیز نتایج به‌دست‌آمده از تطابق با قوانین ساختاری به بردار ویژگی استخراج‌شده از نمونه‌ها افزوده می‌شود. در این راستا ابتدا بردار ویژگی دودویی که از تک-گرام، سرواژه و برچسب نقش گفتاری^{۱۵} (POS) سرواژه تشکیل شده است از نمونه‌های برچسب‌دار به دست می‌آید. از سوی دیگر هر نمونه‌ی دارای برچسب با استفاده از مجموعه‌ای از قوانین مبتنی بر ساختار جملات، تحلیل نحوی می‌شود. این امکان وجود دارد که هر پرسش با بیش از یک قانون تطابق یابد. به این ترتیب انواع پاسخ‌های مورد انتظار پرسش، با استفاده از قوانین تعیین می‌شوند. از این‌رو بردار دودویی دیگری با طول برابر با تعداد انواع پاسخ‌ها، به بردار ویژگی اولیه افزوده می‌شود. یعنی ترکیب این دو بردار به‌عنوان ورودی الگوریتم یادگیری ماشین ارسال می‌شود. این امر سبب می‌شود تا صحت دسته‌بندی در یادگیری افزایش یابد. تاکنون در پژوهش‌هایی که در این راستا صورت پذیرفته است [۶، ۹] ساختار طبقه‌بندی پرسش‌ها در قوانین، مشابه ساختار موردنیاز در دسته‌بندی بوده است. ولی نوآوری اصلی ارائه‌شده در این مقاله، ترکیب بردارهای ویژگی نمونه‌ها با بردارهای حاصل از تطابق با قوانینی با ساختار طبقه‌بندی متفاوت است که به این ترتیب می‌توانند در غنی‌سازی بردارهای ویژگی مشارکت داشته باشند.

همچنین هرچند تاکنون از SVM در پژوهش‌های مختلفی در این حوزه استفاده شده است ولی تاکنون از توسعه‌های جدیدی که در این فرضیه یادگیری داده شده در این حوزه از پژوهش استفاده نشده است. نوآوری جنبی ارائه‌شده در این مقاله، افزایش صحت دسته‌بندی با ماشین بردار پشتیبان دوقلو (Twin SVM) است.

در ادامه این مقاله ابتدا در بخش ۲ مروری بر سایر پژوهش‌های انجام‌شده در حوزه دسته‌بندی پرسش با استفاده از رویکردهای مبتنی بر قانون و مبتنی بر یادگیری ماشین خواهیم داشت. سپس در بخش ۳

سپس خروجی این دسته‌بند، به بردار ویژگی پرسش‌ها اضافه شده و مجدد در ریزدانه‌ها دسته‌بندی می‌شوند.

در [۱۰] الگوریتم SVM خطی برای دسته‌بندی پرسش‌ها مورد استفاده قرار گرفته است. محققان در ارزیابی‌های این پژوهش، بردار ویژگی را متشکل از کیسه کلمات، کیسه‌ای از دنباله‌های نتایی از کلمات و موجودیت‌های نامدار پرسش در نظر گرفته‌اند. آن‌ها ابعاد فضای ویژگی را با استفاده از الگوریتم تفکیک مقادیر منحصربه‌فرد^{۱۷} (SVD) کاهش داده‌اند. نتایج آن‌ها حاکی از این است که استفاده از این الگوریتم، صحت دسته‌بندی را کاهش می‌دهد.

در پژوهش [۱۱] عملکرد چندین تکنیک یادگیری ماشین مانند SVM با هسته درختی، k- نزدیک‌ترین همسایه (KNN)، DT و NB با بهره‌مندی از ویژگی‌های سطحی متن چون کیسه‌ای از کلمات و کیسه‌ای از دنباله‌های نتایی مورد ارزیابی قرار گرفته‌اند. محققان این پژوهش بر اساس نتایج خود اظهار داشته‌اند که استفاده از هسته درختی سبب می‌شود اطلاعات ساختار پرسش نیز به نوعی در فرآیند دسته‌بندی شرکت داشته باشند و این مسئله موجب بهبود عملکرد SVM در فضاهای با ابعاد زیاد می‌گردد.

محققان در [۴] به مقایسه عملکرد مدل‌های NB، KNN و SVM خطی و غیرخطی پرداخته‌اند. در این پژوهش به منظور آموزش مدل یادگیری، از ویژگی‌های مختلف لغوی، نحوی و معنایی پرسش استفاده کرده‌اند. سپس ترکیب‌های مختلف این ویژگی‌ها را به کار گرفته و نتایج حاصل را مقایسه نموده‌اند. نتایج ارزیابی‌های آن‌ها حاکی از این است که مدل یادگیری SVM خطی، بهترین مدل برای دسته‌بندی پرسش‌ها است. همچنین مجموعه‌ای از ویژگی‌ها که بیش‌ترین توانایی را در تفکیک دسته‌ها دارد از تک-گرام‌ها، شکل کلمات، سرواژه پرسش، ویژگی‌های معنایی سرواژه و دسته پرسش تشکیل شده است. شکل کلمات پرسش مبنی بر بزرگ یا کوچک بودن حروف کلمات در پرسش است و دسته پرسش یعنی دسته‌ای که بر اساس سلسله‌مراتب وردنت بیش‌ترین شباهت را به سرواژه پرسش دارد.

جمعی از پژوهشگران در [۷] یکی از مهم‌ترین چالش‌های دسته‌بندی پرسش را بالا بودن ابعاد بردار ویژگی دانسته‌اند. آن‌ها تمرکز خود را بر انتخاب ویژگی‌های معنایی جدید با استفاده از وردنت، بهبود استخراج سرواژه پرسش و رفع ابهام از کلمات پرسش منعطف کرده‌اند تا بتوانند جایگزینی برای ویژگی‌های لغوی بیابند. همچنین به منظور بهبود سرعت یادگیری و تست، الگوریتم ماشین یادگیری افراطی^{۱۸} (ELM) را برگزیده‌اند. ELM یک نوع الگوریتم شبکه عصبی روبه‌جلو^{۱۹} است. نتایج آن‌ها نشان می‌دهد که این الگوریتم نسبت به SVM دقیق‌تر و نسبت به انتشار روبه‌جلو^{۲۰} سریع‌تر است.

۳- روش پیشنهادی

در این مقاله ترکیب روش‌های مبتنی بر قانون و یادگیری ماشین، با هدف افزایش صحت دسته‌بندی پرسش‌ها ارائه شده است. ساختار

شامل ۹۴ گره است که ۴۷ تای آن‌ها گره‌های برگ می‌باشند. این مجموعه، کامل‌ترین و بزرگ‌ترین مجموعه قوانینی است که تاکنون ساخته شده است [۶].

گروهی از پژوهشگران در [۳] ترکیبی از رویکردهای دسته‌بندی مبتنی بر قانون و تطابق الگو را ارائه نموده‌اند. در سیستم پیشنهادی آن‌ها، ابتدا کلیه پرسش‌های برچسب‌دار به مؤلفه شناسایی نقش‌های گفتاری جمله ارسال می‌شوند تا الگوی ساختاری آن‌ها تعیین گردد. سپس همه الگوها بر اساس کلمه پرسشی دسته‌بندی می‌شوند. بدین ترتیب، برای هر پرسش بدون برچسب ورودی، ابتدا الگوی ساختاری و کلمه تمرکز که حاوی بیش‌ترین اطلاعات در پرسش است، شناسایی می‌شود. بر اساس کلمه تمرکز و نقش گفتاری آن در جمله، با استفاده از مجموعه‌ای از قوانین، دسته درشت‌دانه پرسش تعیین می‌شود. سپس برای تعیین دسته ریزدانه، بر اساس کلمه پرسشی آن، در یکی از دسته‌ها تطابق الگو انجام می‌شود. برچسب مربوط به هر الگویی که تطبیق یابد، به پرسش ورودی تعلق می‌گیرد. آن‌ها در پژوهش‌های خود به این نتیجه رسیدند که رویکرد ترکیبی، صحت دسته‌بندی پرسش و شناسایی کلمه تمرکز را بهبود می‌دهد.

محققان در [۶] نیز بخشی از فرآیند دسته‌بندی پرسش‌ها را با استفاده از مجموعه‌ای از ۶۰ قانون انجام داده‌اند. مجموعه قوانین مورد استفاده در این پژوهش، ساختار طبقه‌بندی یکسانی نسبت به ساختار موجود برای دسته‌بندی پرسش‌ها دارند. از این رو توانسته‌اند صحت دسته‌بندی مبتنی بر قانون را نیز ارزیابی نمایند.

۲-۲- روش‌های مبتنی بر یادگیری ماشین

تحقیقات در حوزه دسته‌بندی پرسش با استفاده از رویکرد مبتنی بر یادگیری ماشین از سال ۲۰۰۲ آغاز شد [۱۳]. در این رویکرد به جای استفاده از مجموعه قوانین از پیش تعریف‌شده، بر اساس جملات پرسشی برچسب‌دار، یک مدل دسته‌بندی آموخته شده و بدین ترتیب با مشکلات رویکرد مبتنی بر قانون مقابله می‌شود. اکثر پژوهش‌هایی که تاکنون از این رویکرد استفاده نموده‌اند در مدل یادگیری و یا ویژگی‌های استخراج‌شده برای یادگیری با یکدیگر متفاوت هستند [۵].

لی و روث در [۱۳] یک رویکرد سلسله‌مراتبی برای دسته‌بندی پرسش‌ها با عنوان SNoW ابداع نمودند. این مدل دارای یک معماری یادگیری چنددسته‌ای است که برای فضاهای ویژگی با ابعاد زیاد کارآمد است. SNoW برای هر دسته، یک تابع خطی می‌آموزد. مزیت این مدل این است که امکان وزن‌دهی به کلاس‌ها را دارد [۵]. آن‌ها مجموعه‌ای از ویژگی‌های لغوی، نحوی و معنایی پرسش مانند برش‌های پرسش، برش دربرگیرنده سرواژه، موجودیت‌های نامدار^{۱۶} و واژه‌نامه‌ای از کلماتی که به لحاظ معنایی با پرسش مرتبط هستند را به عنوان بردار ویژگی برگزیدند [۷، ۱۴، ۱۶]. به‌طورکلی رویکرد آن‌ها از دو دسته‌بند تشکیل شده است. یک دسته‌بند برای درشت‌دانه‌ها اعمال می‌شود.

پرسش‌ها بر اساس جداکننده "فاصله خالی"، به آرایه‌ای از کلمات تشکیل‌دهنده آن‌ها شکسته می‌شوند. سپس آرایه هر پرسش، بررسی شده و چنانچه واژه‌ای در آن باشد که هم‌اکنون در واژه‌نامه موجود نیست، به واژه‌نامه افزوده می‌شود. پس از ساخت واژه‌نامه نوبت به ساخت بردار کلمات برای پرسش‌ها می‌رسد. برای بار دوم، کلیه پرسش‌ها بررسی می‌شوند. به ازای هر کلمه از واژه‌نامه که در پرسش وجود دارد، مقدار متناظر ۱ و در غیر این صورت مقدار متناظر ۰ در بردار ویژگی ثبت می‌شود. به این ترتیب برای هر پرسش، برداری با طول برابر با طول واژه‌نامه ساخته می‌شود.

ویژگی‌های نحوی ویژگی‌هایی هستند که نیازمند تحلیل ساختار گرامری پرسش‌ها می‌باشند. از آن جمله می‌توان به سرواژه پرسش و POS سرواژه اشاره کرد. سرواژه، نشان‌دهنده اطلاعاتی است که پرسش در پی یافتن آن‌ها است. تحقیقات پژوهشگران نشان داده است که سرواژه، یکی از حیاتی‌ترین ویژگی‌های پرسش محسوب می‌شود [۴]. شناسایی سرواژه، نیازمند در اختیار داشتن درخت تجزیه پرسش است. در این پژوهش به منظور یافتن درخت تجزیه، از تجزیه‌گر برکلی استفاده شده است. سپس، درخت تجزیه پرسش با استفاده از مجموعه قوانین کالینز بررسی می‌شود. در واقع رویکرد استفاده شده در این پژوهش برای شناسایی سرواژه، همان رویکرد ارائه شده در [۶] است. با این تفاوت که به جای نسخه تغییر یافته قوانین کالینز، خود قوانین اعمال شده‌اند. به محض یافتن سرواژه، در همان لحظه می‌توان POS آن را نیز از درخت تجزیه شناسایی نمود. پس از شناسایی سرواژه تمام پرسش‌ها، به طور مشابه، واژه‌نامه‌هایی برای سرواژه‌ها و POSها ساخته شده و برای هر پرسش، برداری دودویی از سرواژه‌ها و POSها استخراج می‌شود.

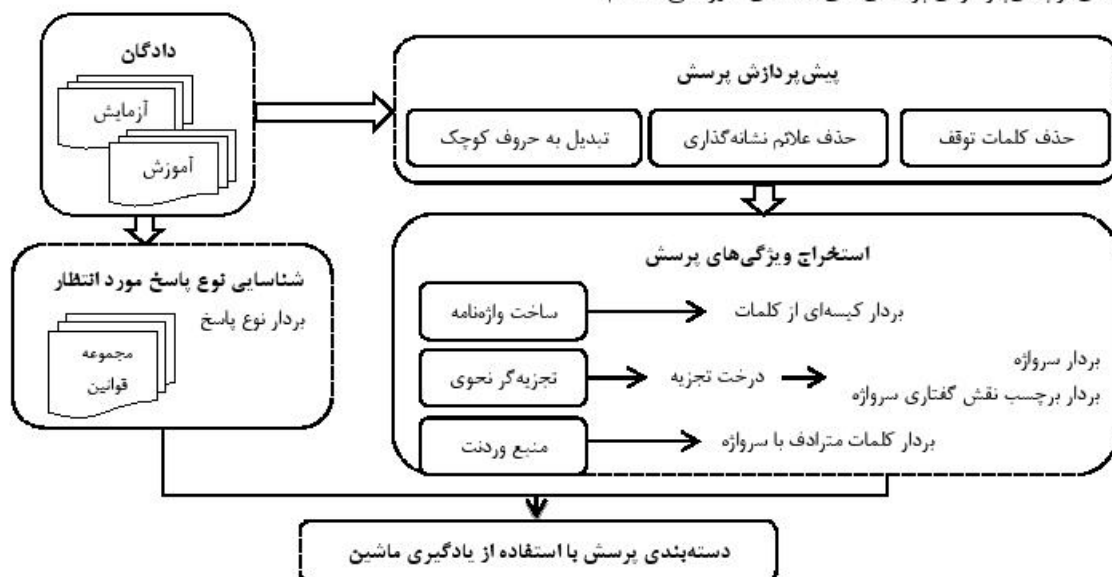
رویکرد پیشنهادی مطابق با شکل ۱ از چهار مؤلفه اصلی (۱) پیش‌پردازش پرسش، (۲) استخراج ویژگی‌های پرسش، (۳) شناسایی نوع پاسخ مورد انتظار پرسش و (۴) دسته‌بندی پرسش، تشکیل شده است.

۳-۱- پیش‌پردازش پرسش

این مؤلفه موظف است کلیه پرسش‌های دادگان آموزشی و آزمایشی را پیش‌پردازش نموده و ساختار نوشتاری تمام پرسش‌ها را تا حد امکان به فرم مشابه یکدیگر تبدیل کند. از آنجا که یک کلمه ممکن است در پرسش‌های متفاوت با صورت‌های نوشتاری مختلف ظاهر شود، لذا در اولین گام، تمام کلمات به حروف کوچک انگلیسی تبدیل می‌شوند. در گام دوم باید علامت‌های نشانه‌گذاری حذف شوند. زیرا علامت نشانه‌گذاری زبان انگلیسی مانند "!"، "?" می‌توانند در برخی پرسش‌ها استفاده شده و در برخی دیگر ظاهر نشوند. سومین گام، حذف کلمات توقف از پرسش است. کلمات توقف کلماتی هستند که به تنهایی بار معنایی ندارند. لازم به تأکید است هر یک از سه گام فوق می‌توانند به صورت اختیاری انجام شوند.

۳-۲- استخراج ویژگی‌های پرسش

میزان موفقیت مدل یادگیری، وابسته به مجموعه ویژگی‌هایی است که از پرسش انتخاب می‌شود. از هر پرسش می‌توان ویژگی‌های مختلف واژگانی، نحوی و معنایی را استخراج نمود. ویژگی‌های واژگانی همان تک-گرامرها یا به بیان دیگر کیسه‌ای از کلمات هستند که مستقیماً از پرسش به دست می‌آیند. به منظور ساخت بردار کیسه کلمات، لازم است ابتدا واژه‌نامه‌ای از تمام کلمات موجود در دادگان آموزشی ساخته شود. به این ترتیب پس از پیش‌پردازش پرسش‌های دادگان آموزشی، تمام



شکل ۱. معماری رویکرد پیشنهادی

ویژگی‌های معنایی بر اساس معنای کلمات پرسش و معمولاً با استفاده از دادگان سومی چون وردنت استخراج می‌شوند. در این پژوهش، ویژگی معنایی مورد استفاده، کلمات مترادف سرواژه پرسش هستند. ممکن است دو پرسش، دارای سرواژه یکسان نباشند اما

ویژگی‌های معنایی بر اساس معنای کلمات پرسش و معمولاً با استفاده از دادگان سومی چون وردنت استخراج می‌شوند. در این پژوهش، ویژگی معنایی مورد استفاده، کلمات مترادف سرواژه پرسش هستند. ممکن است دو پرسش، دارای سرواژه یکسان نباشند اما

قوانین متناظر موجود در این طبقه‌بندی به مجموعه قوانین پژوهش افزوده می‌شوند. سپس پرسش‌هایی که در این طبقه‌بندی هیچ قانونی برای آن‌ها یافت نشده است شناسایی می‌شوند. بیش از ۱۰۰۰ پرسش از دادگان آموزشی، هیچ تطابقی نیافتند. بدین ترتیب، این پرسش‌ها به صورت دستی بررسی شده و بر اساس نوع پاسخ و ساختار آن‌ها، قوانین جدیدی به مجموعه قوانین افزوده می‌شوند. این کار تا زمانی که تمام پرسش‌ها پوشش داده شوند ادامه می‌یابد. تعداد این قوانین به ۱۸۰ عدد می‌رسد که پاسخ موردانتظار پرسش‌ها را در ۵۰ دسته برچسب می‌زنند. انواع پاسخ‌های مورد انتظار مبتنی بر قوانین در جدول ۱ و نمونه‌هایی از قوانین در جدول ۲ قابل مشاهده هستند.

بر خلاف سایر پژوهش‌ها، در رویکرد پیشنهادی این مقاله، برای یک پرسش، تعداد برچسب مبتنی بر قانون محدودیت ندارد. ضمن اینکه مجموعه برچسب‌های مبتنی بر قانون، متفاوت با مجموعه برچسب‌های ریزدانه و درشت‌دانه دادگان است. به عنوان مثال سایر پژوهش‌ها، برای پرسش "How many species of the great white shark are there?" پس از تطبیق با قوانین، فقط یک برچسب "Num:Count" را انتساب می‌دهند. این برچسب متعلق به مجموعه برچسب‌های اصلی دادگان نیز هست. اما در رویکرد این مقاله، به این پرسش، چهار برچسب "NumericalQuantity"، "Quantity"، "TemporalQuantity" و "General Information" انتساب داده می‌شود. یعنی حالت‌های محتمل دیگر را نیز پوشش می‌دهد. پیش‌بینی می‌شود این رویکرد، سبب بهبود فرآیند یادگیری و شناسایی هرچه صحیح‌تر پرسش‌های مشابه گردد.

سرواژه‌های مترادف با یکدیگر داشته باشند. لذا پیش‌بینی می‌شود که استخراج کلمات مترادف با هر سرواژه می‌تواند فاصله معنایی پرسش‌ها از یکدیگر را کاهش دهد. به منظور یافتن کلمات مترادف از رویکرد ارائه شده در [۶] استفاده شده است. با توجه به زیاد بودن تعداد کلمات مترادف، ساخت واژه‌نامه برای این کلمات، امری زمان‌بر است و حجم بردارها را به شدت افزایش می‌دهد. لذا در این پژوهش برای ساخت واژه‌نامه، یکبار کلمات مترادف با حداقل دو بار مشاهده در دادگان و بار دیگر کلمات با حداقل ۳ بار تکرار در دادگان انتخاب شده‌اند و بر این اساس بردارهای دودویی هر پرسش شکل گرفته‌اند.

۳-۳- شناسایی نوع پاسخ مورد انتظار پرسش

همان‌طور که پیش از این در بخش ۲-۱- اشاره شد، بسیاری از پژوهش‌های اولیه در حوزه دسته‌بندی پرسش، از مجموعه‌ای از قوانین دست‌ساز استفاده کرده‌اند. ایراد قوانین دست‌ساز، نامناسب بودن آن‌ها به هنگام تغییر از یک دادگان به دادگان دیگر و زمان‌بر بودن ساخت مجموعه‌ای کامل از قوانین است. در این پژوهش پیشنهاد شده است که هر پرسش ورودی، ابتدا به مجموعه‌ای از قوانین ساختاری داده شود تا نوع پاسخ مورد انتظار آن پیش‌بینی شود. یک پرسش ممکن است با بیش از یک قانون تطابق یابد. پس از آن برای هر پرسش، برداری به طول تعداد انواع پاسخ‌ها در نظر گرفته می‌شود. درایه متناظر با نوع پاسخ‌های مورد انتظار پیش‌بینی شده برای هر پرسش در بردار دودویی مقدار ۱ و مابقی درایه‌ها مقدار ۰ می‌گیرند.

پیاده‌سازی این مؤلفه، نیازمند ساخت مجموعه‌ای پوشا از قوانین است. برای این منظور، از طبقه‌بندی ویکلوپدیا [۸] به عنوان پایه کار بهره گرفته شده است. ابتدا برای تمام پرسش‌های دادگان آموزشی،

جدول ۱: ساختار طبقه‌بندی مبتنی بر قوانین ارائه شده در این پژوهش

نوع پاسخ مورد انتظار
Yes/No - True/False
food - Substance - BodyPart - Liquid
ColorQ - Shape - Adjective - TextQ - Disease
Agent - NameQ - LastName - FirstName - CompanyNameQ - OrganizationName
Person - GroupOfPeople - Animal - OccupationPerson - GeographicalPerson - ProperPlace
Quantity - NumericalQuantity - MonetaryQuantity - TemporalQuantity
- MassQuantity - VolumeQuantity - DistanceQuantity - AreaQuantity - Date
StateDistrict - City - Country - Address - EmailAddress - URL - PhoneNumber
Consequent - WhyFamous - Antecedent - Influence - Bio
GeneralInformation - ExpressionOrigin - Use - Definition
Reason - Description

جدول ۲: نمونه‌هایی از قوانین و عبارات‌های منظم جهت تعیین نوع پاسخ مورد انتظار پرسش

قوانین و عبارات‌های منظم	نوع پاسخ مورد انتظار
^What(is are was would)(.*)?(first girl boy guy)(%.*? .*?)name.*	First Name
(^How(is many much few hot loud cold)) ^At(is what density)	Quantity
(^What(is are was were)(.*)?(odd statistic chance size)) (^What(is is was were are)(.*)?(address apt. number)) ^Where(is do did does)(.*)?live	Address
Describe(is the .*?) (^According(is to is .*?) .*?)(is H)ow(is long)	Description
(^What(is is was were is the .*?)(time life sexpe ctancy age limit)) ^At(is what age)(.*)?^What(is age does)(.*)?at ^How(is old)((.*)?(is was are) how old)	Temporal Quantity
(^How(is much many often)(.*)?(second minute hour day month year decade .*?))	

۳-۴- دسته‌بندی پرسش

توقف، جستجو انجام شود. اگر تعداد کلمات موجود در پرسش نام را با W_i تعریف نماییم، آنگاه تعداد کل کلمات آموزش با توجه به رابطه (۲) محاسبه می‌شود.

$$W = \sum_{i=1}^n W_i \quad (2)$$

از این رو، چنانچه تعداد کلمات توقف تعریف شده برای سیستم، l_1 باشد، در این صورت با استفاده از جستجوی دودویی داریم:

$$T(sW) = W \times (\log_2(l_1)) \quad (3)$$

اگر طول بردار علائم نشانه‌گذاری نیز برابر با l_2 باشد، پیچیدگی زمانی حذف علائم نشانه‌گذاری از رابطه (۴) محاسبه می‌شود.

$$T(m) = W \times (\log_2(l_2)) \quad (4)$$

زمان $T(e)$ به نوع ویژگی‌های استخراج شده از پرسش وابسته است. به بیان دیگر، برابر با مجموع پیچیدگی زمانی لازم برای استخراج ویژگی‌هایی چون کیسه کلمات $T(b)$ ، سرواژه $T(h)$ و برچسب نقش گفتاری سرواژه $T(hp)$ است. ساخت بردار ویژگی کیسه کلمات، بر اساس واژه‌نامه تولید شده در مرحله آموزش انجام می‌شود. در بدترین حالت، اگر کلمات پرسش‌های آموزش، اشتراکی با یکدیگر نداشته باشند، طول واژه‌نامه برابر l_3 خواهد بود. لذا برای ساخت بردار کیسه کلمات برای هر پرسش، باید کلمات آن با کلیدی کلمات واژه‌نامه مقایسه شوند. پیچیدگی زمانی این فرآیند برابر است با:

$$T(b) = l_3 \times W \quad (5)$$

در این پژوهش، سرواژه و برچسب نقش گفتاری سرواژه هر پرسش، همزمان تعیین می‌شوند. به این ترتیب که درخت تجزیه هر پرسش، سطح به سطح مورد بررسی قرار می‌گیرد تا به گره برگی که مطابق قوانین، سرواژه پرسش محسوب می‌شود دست یابیم. لذا پیچیدگی زمانی استخراج این دو ویژگی، وابسته به زمان استخراج درخت تجزیه پرسش با استفاده از تجزیه‌گر برکلی $T(parse)$ و زمان بررسی سطوح درخت با استفاده از قوانین کالینز $T(collins)$ است. پس از تعیین آن‌ها، چنانچه طول واژه‌نامه هر یک به ترتیب برابر با l_4 و l_5 باشد بردار ویژگی مربوط به هر یک تولید می‌شود. بنابراین داریم:

$$+l_4 + l_5 T(h) + T(hp) = \sum_{i=1}^n (T_i(parse) + T_i(collins)) \quad (6)$$

به منظور استخراج نوع پاسخ مورد انتظار هر پرسش، باید آن را با مجموعه ۶۰ قانون ارزیابی نموده و درایه متناظر با نوع پاسخ، در بردار مربوطه مقدار ۱ بگیرد. لذا داریم:

$$T(r) = 60n \approx n \quad (7)$$

زمان دسته‌بندی پرسش یا همان $T(c)$ وابسته به نوع الگوریتم یادگیری ماشین انتخاب شده است. چنانچه تعداد دسته‌ها برابر با c باشد آنگاه در صورت انجام آموزش به صورت یک به همه، تعداد مدل‌های یادگیری برابر c خواهد بود. در نتیجه، هر نمونه بدون برچسب،

در مؤلفه دسته‌بندی پرسش، نمونه‌های آموزشی در اختیار الگوریتم یادگیری قرار می‌گیرند تا بر اساس آن‌ها مدلی را برای دسته‌بندی بیاموزد. هر یک از نمونه‌های آموزشی دارای یک برچسب درشت‌دانه و یک برچسب ریزدانه است. لذا دسته‌بندی پرسش باید در دو سطح انجام پذیرد. ابتدا با بهره‌گیری از ترکیب‌های مختلف بردارهای ویژگی دادگان آموزشی، آموزش مدل یادگیری در ۶ دسته درشت‌دانه انجام می‌شود. سپس نمونه آزمایشی با استفاده از مدل‌های آموخته‌شده بررسی می‌شود و دسته درشت‌دانه آن تعیین می‌گردد. برای تعیین دسته ریزدانه می‌توان به دو شیوه مختلف عمل کرد:

الف) آموزش مدل یادگیری بدون توجه به دسته درشت‌دانه و بر اساس تمام نمونه‌های آموزشی انجام می‌شود. سپس برچسب ریزدانه پرسش آزمایشی تعیین می‌شود.

ب) پس از تعیین دسته درشت‌دانه پرسش، آموزش مدل یادگیری تنها روی نمونه‌های آموزشی که در همان دسته هستند، انجام می‌شود. در این صورت، دسته‌ای که برای ریزدانه در نظر گرفته می‌شود، حتما زیردسته‌ای از دسته درشت‌دانه خواهد بود.

از آنجا که نتایج ارزیابی‌های مختلف پژوهشگران نشان داده است که موفق‌ترین الگوریتم برای دسته‌بندی پرسش SVM است [۴]. این پژوهش نیز SVM و نسخه Twin SVM را به عنوان مدل یادگیری خود برگزیده است. الگوریتم SVM را می‌توان در دو حالت یک به همه یا یک به یک انجام داد. ضمن اینکه چنانچه برای یک پرسش، دو دسته درشت‌دانه، امتیاز برابری کسب نمایند، دسته‌ای انتخاب می‌شود که احتمال اولیه بیش‌تری داشته باشد. مقادیر احتمالات اولیه برای هر دسته، بر اساس نمونه‌های آموزشی تعیین می‌شوند.

۳-۵- تحلیل زمانی

یکی از مهم‌ترین عوامل در انتخاب مؤثر ویژگی‌ها و مدل یادگیری، ارزیابی زمان است. از آنجا که آموزش مدل یادگیری بر اساس نمونه‌های برچسب‌دار، به صورت برون خط صورت می‌گیرد، لذا آنچه در ارزیابی رویکردهای مختلف حائز اهمیت است، زمان مورد نیاز برای دسته‌بندی پرسش‌های بدون برچسب است. پیچیدگی زمانی این فرآیند یعنی $T(test)$ برابر با مجموع پیچیدگی زمانی فازهای پیش‌پردازش $T(p)$ ، استخراج ویژگی $T(e)$ ، شناسایی نوع پاسخ مورد انتظار $T(r)$ و دسته‌بندی با رویکرد یادگیری ماشین $T(c)$ است. چنانچه تعداد نمونه‌های بدون برچسب برابر با n باشد داریم:

$$T(test) = \sum_{i=1}^n (T_i(p) + T_i(e) + T_i(r) + T_i(c)) \quad (1)$$

پیچیدگی $T(p)$ در رابطه (۱)، می‌تواند متشکل از پیچیدگی زمانی حذف کلیه کلمات توقف $T(sw)$ یا زمان حذف کلیه علائم نشانه‌گذاری $T(m)$ و یا هر دو باشد. به منظور حذف کلمات توقف باید به ازای تک تک کلمات موجود در هر پرسش، در میان درایه‌های آرایه‌ای شامل کلمات

برای آموزش ۰/۰۴ ثانیه به طول انجامید. فرآیند ساخت واژه‌نامه به مدت ۲/۵۴ ثانیه، استخراج بردار کیسه کلمات بدون کلمات توقف و علائم نشانه‌گذاری برای کلیه نمونه‌های آموزش به مدت ۱۳/۳۳ ثانیه و برای همه نمونه‌های تست ۱/۴۷ ثانیه، استخراج سرواژه نمونه‌های آموزش ۱۶۵/۱۹ ثانیه و برای نمونه‌های تست ۱۸/۴۹ ثانیه به طول انجامید. استخراج ویژگی مبتنی بر قانون برای نمونه‌های آموزش ۱۳:۲۸/۷۹ دقیقه و برای نمونه‌های تست ۲۹/۲۲ ثانیه زمان صرف نمود.

جدول ۴: طول بردار هر یک از ویژگی‌های استخراج شده از پرسش

طول بردار ویژگی	ویژگی
۸۴۳۷	تک-گرام بدون کلمات توقف و علائم نشانه‌گذاری (پس از تبدیل به حروف کوچک)
۲۲۶	کلمات توقف (پس از تبدیل به حروف کوچک)
۸	علائم نشانه‌گذاری
۹۱۱۳	تک-گرام بدون کلمات توقف و علائم نشانه‌گذاری (پیش از تبدیل به حروف کوچک)
۳۱۹	کلمات توقف (پیش از تبدیل به حروف کوچک)
۲۵۱۰	سرواژه
۲۳	برچسب نقش گفتاری سرواژه
۵۰	ویژگی مبتنی بر قانون (نوع پاسخ مورد انتظار)
۱۱۱۲۵	مترادف‌های سرواژه
۵۹۷۹	مترادف‌های سرواژه که حداقل دو بار تکرار شده‌اند
۴۱۰۵	مترادف‌های سرواژه که حداقل سه بار تکرار شده‌اند

در دومین گام، پس از استخراج ویژگی‌های مختلف پرسش و بردارهای دودویی متناظر با آن‌ها، بردار تک-گرام بدون کلمات توقف و علائم نشانه‌گذاری را به عنوان پارامتر ورودی به الگوریتم SVM متلب با کرنل خطی داده‌ایم. صحت پایه برای دسته‌های درشت‌دانه، برابر با ۵۵/۶٪ به دست آمده است. سپس به منظور بررسی نقش بردار کلمات مترادف سرواژه در صحت دسته‌بندی درشت‌دانه‌ها، آن را به بردار تک-گرام الحاق نموده، به الگوریتم SVM داده و صحت را مورد ارزیابی قرار داده‌ایم. مطابق جدول ۵، نتایج دسته‌بندی با استفاده از بردار کلمات مترادفی که حداقل دو بار تکرار شده‌اند، نشان‌دهنده کاهش صحت به میزان ۲/۱٪ و لذا ناکارآمد بودن این ویژگی در بهبود صحت است. استفاده از بردار مترادف‌ها با حداقل سه بار تکرار نیز با کاهش صحت به میزان ۱٪ همراه بوده است. از این رو در مراحل بعدی این پژوهش از استفاده از آن‌ها صرف نظر نموده‌ایم.

جدول ۵: صحت دسته‌بندی برای درشت‌دانه‌ها با استفاده از بردار کلمات

مترادف سرواژه در SVM یک به همه

بدون ویژگی مبتنی بر قانون	بردار ویژگی
۷۵۵۶ (۲۷۸ / ۵۰۰)	تک-گرام
۷۵۱۸ (۲۵۹ / ۵۰۰)	تک-گرام + کلمات مترادف سرواژه (مترادف‌هایی که حداقل دو بار تکرار شده‌اند)
۷۵۰۱ (۲۵۴ / ۵۰۰)	تک-گرام + کلمات مترادف سرواژه (مترادف‌هایی که حداقل سه بار تکرار شده‌اند)

بردار کیسه کلمات می‌تواند فاقد کلمات توقف و علائم نشانه‌گذاری باشد. شامل آن‌ها باشد. ضمن اینکه می‌تواند پیش از تبدیل کلیه کلمات به

باید به c مدل یادگیری داده شده و دسته آن تعیین شود T(predict). بنابراین مطابق رابطه (۸) داریم:

$$T(c) = n \times \sum_{k=1}^c T_k(\text{predict}) \quad (8)$$

مطابق توضیحات ارائه شده، زمان موردنیاز برای ارزیابی، برابر با میانگین زمان لازم برای همه نمونه‌های بدون برچسب است. یعنی:

$$\text{Time for Evaluation} = T(\text{test}) / n \quad (9)$$

۴- نتایج تجربی آزمایش‌ها

دادگان UIUC یکی از شناخته شده ترین دادگان در حوزه دسته‌بندی پرسش است که در سال ۲۰۰۲ تدوین شده است [۲۰]. UIUC از ۵۵۰۰ نمونه آموزشی و ۵۰۰ نمونه تست که از ۴ منبع مختلف گردآوری شده‌اند [۱۶]، تشکیل شده است. هر نمونه دارای یک برچسب درشت‌دانه و یک برچسب ریزدانه است [۵، ۷، ۱۳]. هر پرسش در یک خط تعریف شده و ساختار نمونه‌ها مطابق زیر است:

<پرسش> <فاصله خالی> <برچسب ریزدانه> <برچسب درشت‌دانه>

تعداد کل دسته‌های درشت‌دانه شش دسته و تعداد کل دسته‌های ریزدانه، ۵۰ دسته مطابق با جدول ۳ است. این دادگان برای استفاده عموم در محیط وب قرار داده شده و قابل دسترسی است [۲۰]. از آنجاکه تمام پژوهش‌هایی که طبقه‌بندی دولایه پیشنهاد شده توسط لی و روث را انتخاب کرده‌اند از این دادگان استفاده نموده‌اند [۸-۱۱، ۱۲، ۱۴ و ۱۵]، در این پژوهش نیز ارزیابی‌ها و آزمایش‌های خود را روی این داده‌ها انجام داده‌ایم تا نتایج کار قابل مقایسه با سایر پژوهش‌ها باشد.

جدول ۳: طبقه‌بندی دولایه برای دسته‌بندی پرسش‌ها همراه با توزیع

دسته‌های درشت‌دانه در دادگان UIUC [۵، ۱۵]

ریزدانه	درشت‌دانه
Abbreviation - Expansion	ABBREVIATION (۸۶)
Animal - Body - Color - Creative - Currency - Disease - Event - Food - Instrument - Lang - Letter - Other - Plant - Product - Religion - Sport - Substance - Symbol - Technique - Term - Vehicle - Word	ENTITY (۱۲۵۰)
Definition - Description - Manner - Reason	DESCRIPTION (۱۱۶۲)
Group - Indicator - Title - Description	HUMAN (۱۲۲۳)
City - Country - Mountain - Other - State	LOCATION (۸۳۵)
Code - Count - Date - Distance - Money - Order - Other - Period - Percent - Speed - Temp - Size - Weight	NUMERIC (۸۹۶)

ما آزمایش‌ها را در شش گام اصلی، مرحله به مرحله انجام داده‌ایم. در گام اول، پرسش‌های دادگان را یک‌به‌یک خوانده و ضمن جداسازی برچسب‌ها از نمونه‌ها و پیش‌پردازش آن‌ها، برداری متناظر با هر نوع ویژگی به صورت جداگانه استخراج نموده‌ایم. طول هر یک از این بردارها مطابق با جدول ۴ است. حذف کلمات توقف برای همه نمونه‌های آموزش ۰/۶۳ ثانیه و برای نمونه‌های تست ۰/۰۳ ثانیه زمان برد. حذف علائم نشانه‌گذاری نیز برای نمونه‌های تست چندان زمان بر نبوده و

در پنجمین گام، ترکیب‌های مختلف ویژگی‌های استخراج شده از پرسش را در دو الگوریتم SVM و Twin SVM مورد ارزیابی قرار داده ایم. همان‌طور که در جدول ۸ ملاحظه می‌شود، در پیش‌بینی دسته درشت‌دانه، بهترین صحت بدون استفاده از بردار ویژگی مبتنی بر قانون، متعلق به زمانی است که بردار ویژگی، متشکل از کیسه کلمات، سرواژه پرسش و برچسب نقش گفتاری سرواژه پرسش باشد که در آن به صحت ۷۹/۸٪ دست یافته‌ایم. چنانچه همین دسته‌بندی را بار دیگر پس از الحاق با بردار ویژگی مبتنی بر قانون انجام دهیم نتیجه ۸۷/۴٪ حاصل می‌شود. این نتیجه، مؤثر بودن رویکرد پیشنهادی این پژوهش مبنی بر ترکیب دو رویکرد مبتنی بر قانون و یادگیری را تأیید می‌کند. از سوی دیگر، اگر همین آزمایش‌ها را با الگوریتم Twin SVM انجام دهیم صحت دسته‌بندی، پیش و پس از استفاده از بردار مبتنی بر قانون به ترتیب برابر با ۸۸/۴ و ۹۱ درصد می‌شود. این امر نشان‌دهنده عملکرد بهتر Twin SVM نسبت به SVM در دسته‌بندی درشت‌دانه پرسش است.

همچنین علاوه بر ارزیابی دقت، اقدام به اندازه‌گیری میانگین زمان دسته‌بندی برای یک نمونه تست در دو الگوریتم SVM و Twin SVM نموده ایم. این اندازه‌گیری‌ها با استفاده از توابع Tic و Toc متلب و به ازای بردارهای ویژگی مختلف انجام شده‌اند که نتایج آن در جدول ۸ برحسب میلی‌ثانیه آمده است. همان‌طور که ملاحظه می‌شود الگوریتم Twin SVM برای دسته‌بندی هر نمونه تست سریع‌تر از الگوریتم SVM عمل می‌نماید.

به‌منظور پیش‌بینی دسته ریزدانه پرسش، عملیات دسته‌بندی تنها به ازای ریزدانه‌های دسته درشت‌دانه‌ای که برای آن پیش‌بینی شده است انجام می‌شود. اگر دسته درشت‌دانه به درستی تعیین نشود، تلاش برای تعیین دسته ریزدانه بی‌ثمر خواهد بود. از این‌رو لازم به تأکید است که صحت دسته‌بندی درشت‌دانه، عملکرد دسته‌بندی ریزدانه را تحت تأثیر قرار می‌دهد. نتایج دسته‌بندی ریزدانه در جدول ۹ حاکی از آن است که ترکیب ویژگی مبتنی بر قانون با بهترین بردار ویژگی، سبب بهبود صحت SVM به میزان ۶/۶٪ (۷۰/۱۶٪ به ۷۷/۲٪) شده است. این در حالی است که Twin SVM صحت دسته‌بندی را با بهره‌گیری از ویژگی مبتنی بر قانون به ۸۰٪ رسانده است. بنابراین هر دو رویکرد پیشنهادی صحت دسته‌بندی ریزدانه‌ها را نیز افزایش می‌دهند.

حروف کوچک یا پس از تبدیل ایجاد شود. لذا در گام سوم به‌منظور بررسی تأثیر وجود یا عدم وجود بردار کلمات توقف و علائم نشانه‌گذاری در صحت دسته‌بندی پرسش، پیش از تبدیل پرسش‌ها به حروف کوچک، الگوریتم را به ازای حالت‌های مختلف ارزیابی نموده‌ایم. همان‌طور که در جدول ۶ مشخص است، برخلاف آنچه در سایر مقالات اشاره شده، صحت دسته‌بند در صورت باقی ماندن کلمات توقف، به هنگام عدم استفاده از بردار مبتنی بر قوانین، به میزان ۲۳٪ و به هنگام استفاده از آن به میزان ۸۸٪ بیش‌تر از زمانی است که کلمات توقف را حذف نماییم. از سوی دیگر، باقی ماندن علائم نشانه‌گذاری در پرسش، صحت دسته‌بندی را نسبت به زمانی که فقط کلمات توقف در بردار موجودند به میزان ۰/۸ تا ۱ درصد کاهش می‌دهد. لذا پس از این، مبنای کار خود را عدم حذف کلمات توقف از پرسش گذاشته‌ایم؛ از این‌رو منظور ما از کیسه کلمات (BOW)، کلمات استخراج‌شده از پرسش به همراه کلمات توقف خواهد بود.

جدول ۶: صحت دسته‌بند برای درشت‌دانه‌ها پس از حذف کلمات توقف و

علائم نشانه‌گذاری در SVM یک به همه

بردار ویژگی	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون
تک-گرام	۷۵۵/۶ (۲۷۸ / ۵۰۰)	۳۸۹ / ۵۰۰
تک-گرام + کلمات توقف	۷۷۸/۶ (۳۹۳ / ۵۰۰)	۷۷۷/۸
تک-گرام + کلمات توقف + علائم نشانه‌گذاری	۷۷۷/۸ (۳۸۹ / ۵۰۰)	۷۸۶/۶

از سوی دیگر، مطابق آنچه در جدول ۴ ارائه شده است، طول بردارهای ویژگی پیش از تبدیل به حروف کوچک، ۶۷۶ واحد بیش‌تر از طول بردارها پس از تبدیل است. بدیهی است هرچه طول بردارها بیش‌تر باشد، حجم محاسبات لازم در مدل یادگیری نیز بیش‌تر خواهد بود. اما برای اطمینان از تأثیر مثبت این تبدیل در صحت دسته‌بندی، در گام چهارم، آزمایش‌های خود را در هر دو حالت انجام داده‌ایم که نتایج آن در جدول ۷ آمده است. آنچه از جدول ۷ برداشت می‌شود این است که تبدیل کلمات به صورت نوشتاری با حروف کوچک می‌تواند سبب بهبود صحت دسته‌بندی پرسش‌ها با استفاده از SVM شود. از این‌رو در ادامه پژوهش، منظور ما از بردار کیسه کلمات، برداری است که از تک-گرام‌ها همراه با کلمات توقف پس از تبدیل به حروف کوچک تشکیل شده است.

جدول ۷: صحت (۶) دسته‌بندی پرسش‌ها با استفاده از SVM یک به همه پیش از تبدیل کلمات به حروف کوچک (سمت راست) و پس از تبدیل (سمت چپ)

دسته‌های درشت‌دانه				بردار ویژگی
پس از تبدیل به حروف کوچک		پیش از تبدیل به حروف کوچک		
با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	
۷۸۲/۲ (۴۱۶ / ۵۰۰)	۷۸۱/۶ (۳۹۳ / ۵۰۰)	۷۸۶/۶ (۴۳۳ / ۵۰۰)	۷۸۱/۶ (۳۹۳ / ۵۰۰)	کیسه کلمات
۷۸۶ (۴۳۰ / ۵۰۰)	۷۸۰ (۴۰۰ / ۵۰۰)	۷۸۵/۶ (۴۲۸ / ۵۰۰)	۷۸۰/۸ (۴۰۴ / ۵۰۰)	کیسه کلمات + سرواژه
۷۸۴/۳ (۴۲۱ / ۵۰۰)	۷۸۲/۳ (۳۹۱ / ۵۰۰)	۷۸۶/۶ (۴۳۳ / ۵۰۰)	۷۸۱/۴ (۳۹۲ / ۵۰۰)	کیسه کلمات + برچسب نقش گفتاری سرواژه
۷۸۷/۴ (۴۳۷ / ۵۰۰)	۷۷۹/۸ (۳۹۹ / ۵۰۰)	۷۸۶/۳ (۴۳۱ / ۵۰۰)	۷۸۲/۳ (۴۱۱ / ۵۰۰)	کیسه کلمات + سرواژه + برچسب نقش گفتاری سرواژه

جدول ۸: میانگین زمان دسته‌بندی هر نمونه تست با استفاده از الگوریتم‌های SVM و Twin SVM

درشت دانه				بردار ویژگی
Twin SVM		SVM		
با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	
۰.۰۶۷۶۶	۰.۰۷۲۸	۰.۲۵۰۶۷	۰.۴۰۸۴۱	کیسه کلمات
۰.۰۸۹۴۹	۰.۰۸۲۵	۰.۱۶۰۲۸	۰.۲۰۴۷۳	کیسه کلمات + سروازه
۰.۱۵۵۲۶	۰.۰۶۶۹۱	۰.۱۲۲۰۴	۰.۱۱۵۹۳۰	کیسه کلمات + برجسب نقش گفتاری سروازه
۰.۱۷۰۱۵۲	۰.۲۲۹۲۹	۶۸	۱۱۵۱۷۰۷	کیسه کلمات + سروازه + برجسب نقش گفتاری سروازه

جدول ۹. نتایج اجرای دو الگوریتم SVM و TWSVM یک به همه، با ویژگی‌های مختلف یک‌بار بدون ویژگی مبتنی بر قانون و بار دیگر با آن

ریزدانه				درشت‌دانه				بردار ویژگی
Twin SVM		SVM		Twin SVM		SVM		
با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	با ویژگی مبتنی بر قانون	بدون ویژگی مبتنی بر قانون	
۱/۷۷۱۶ (۳۸۸/۵۰۰)	۱/۷۳ (۳۶۵/۵۰۰)	۱/۷۲۱۶ (۳۶۳/۵۰۰)	۱/۶۵۱۶ (۳۲۸/۵۰۰)	۱/۸۸ (۴۴/۵۰۰)	۱/۸۲۱۶ (۴۱۳/۵۰۰)	۱/۸۳۱۲ (۴۱۶/۵۰۰)	۱/۷۸۱۶ (۳۹۳/۵۰۰)	کیسه کلمات
۱/۵۳۱۴ (۲۶۷/۵۰۰)	۱/۲۹۱۲ (۱۹۶/۵۰۰)	۱/۵۲۱۶ (۲۶۳/۵۰۰)		۱/۶۸۱۸ (۳۴۴/۵۰۰)	۱/۴۹۱۲ (۲۴۶/۵۰۰)	۱/۶۰۱۲ (۳۰۱/۵۰۰)	همگرا نمی‌شود	سروازه
۱/۴۸ (۲۴۰/۵۰۰)	۱/۴۱۶ (۲۳/۵۰۰)	۱/۸۱۲ (۴۱/۵۰۰)	۱/۴۱۴ (۲۲/۵۰۰)	۱/۶۲ (۳۱/۵۰۰)	۱/۲۳۱۴ (۱۱۷/۵۰۰)	۱/۲۳۱۶ (۱۱۸/۵۰۰)	۱/۲۳۱۲ (۱۱۶/۵۰۰)	برجسب نقش گفتاری سروازه
۱/۵۴۱۶ (۲۷۳/۵۰۰)	۱/۴۱۱۶ (۲۰۸/۵۰۰)	۱/۵۴۱۴ (۲۷۲/۵۰۰)	-	۱/۶۸۱۲ (۳۴۱/۵۰۰)	۱/۵۲۱۶ (۲۶۳/۵۰۰)	۱/۶۳۱۲ (۳۱۶/۵۰۰)	همگرا نمی‌شود	سروازه + برجسب نقش گفتاری سروازه
۱/۷۸۱۸ (۳۹۴/۵۰۰)	۱/۷۴۱۴ (۳۷۲/۵۰۰)	۱/۷۶۱۶ (۳۸۳/۵۰۰)	۱/۶۹۱۲ (۳۴۶/۵۰۰)	۱/۹۰۱۸ (۴۵۴/۵۰۰)	۱/۸۵۱۲ (۴۲۶/۵۰۰)	۱/۸۶ (۴۳۰/۵۰۰)	۱/۸۰ (۴۰۰/۵۰۰)	کیسه کلمات + سروازه
۱/۷۷۱۶ (۳۸۸/۵۰۰)	۱/۷۳۱۶ (۳۶۸/۵۰۰)	۱/۷۴۱۴ (۳۷۲/۵۰۰)	۱/۶۵۱۶ (۳۲۸/۵۰۰)	۱/۸۸۱۲ (۴۴۱/۵۰۰)	۱/۸۳۱۴ (۴۱۷/۵۰۰)	۱/۸۴۱۲ (۴۲۱/۵۰۰)	۱/۷۸۱۲ (۳۹۱/۵۰۰)	کیسه کلمات + برجسب نقش گفتاری سروازه
۱/۸۰ (۴۰۰/۵۰۰)	۱/۷۸۱۴ (۳۹۲/۵۰۰)	۱/۷۷۱۲ (۳۸۶/۵۰۰)	۱/۷۰۱۶ (۳۵۳/۵۰۰)	۱/۹۱ (۴۵۵/۵۰۰)	۱/۸۸۱۴ (۴۴۲/۵۰۰)	۱/۷۸۱۴ (۴۳۷/۵۰۰)	۱/۷۹۱۸ (۳۹۹/۵۰۰)	کیسه کلمات + سروازه + برجسب نقش گفتاری

جدول ۱۱ ارائه شده‌اند. صحت پایه این پژوهش به ازای اجرای الگوریتم SVM متلب با استفاده از بردار ویژگی کیسه کلمات به میزان ۱/۷۸۱۶ است درحالی‌که اجرای همین رویکرد در پژوهش [۶] با استفاده از SVMf متلب با کرنل خطی و با مقدار پیش‌فرض $C=1$ ، به صحت ۱/۸۸ دست یافته است. متأسفانه در [۶] هیچ صحبتی پیرامون ابعاد دقیق بردارهای ویژگی نشده است که بتوان بر اساس آن، تفاوت بردارهای پژوهش جاری را ارزیابی نمود. اما با این وجود رویکرد پیشنهادی این مقاله، بدون استفاده از ویژگی‌های معنایی مانند کلمات مترادف که با کمک منبع سومی چون وردنت استخراج می‌شوند، تنها با استفاده از ویژگی‌های نحوی و واژگانی توانسته است صحتی بیش‌تر از متوسط صحت پژوهش‌های اخیر را کسب نماید. پیش‌بینی می‌شود که چنانچه بتوان صحت پایه را به مرزهای دانش رساند، صحت رویکرد پیشنهادی به میزان قابل توجهی افزایش یابد.

تفاوت صحت دسته‌بندی ریزدانه‌ها نیز به دلیل رویکرد اتخاذشده در این پژوهش مبنی بر دسته‌بندی پرسش‌ها فقط به ازای ریزدانه‌های فرزند درشت‌دانه پیش‌بینی شده برای پرسش است. زیرا صحت دسته‌بندی درشت‌دانه‌ها، عملکرد دسته‌بندی ریزدانه‌ها را تحت تأثیر

در گام ششم علاوه بر Twin SVM نسخه‌های دیگری چون STPM SVM و WLT SVM را نیز مورد آزمایش قرار دادیم. همان‌طور که در جدول ۱۰ مشاهده می‌شود، نتایج این آزمایش‌ها حاکی از آن است که نسخه Twin SVM همچنان به‌عنوان قوی‌ترین نسخه SVM در دسته‌بندی پرسش عمل می‌کند.

جدول ۱۰: مقایسه عملکرد نسخه‌های مختلف SVM برای درشت‌دانه‌ها در حالت یک به همه

WLT SVM	STPM SVM	Twin SVM	بردار ویژگی
۱/۲۱۱۴ (۱۰۷/۵۰۰)	۱/۵۱۱۶ (۲۵۸/۵۰۰)	۱/۸۲۱۶ (۴۱۳/۵۰۰)	کیسه کلمات
۱/۲۱۱۲ (۱۰۶/۵۰۰)	۱/۵۲۱۶ (۲۶۳/۵۰۰)	۱/۸۸ (۴۴۰/۵۰۰)	کیسه کلمات + ویژگی مبتنی بر قانون
۱/۱۹۱۶ (۹۸/۵۰۰)	۱/۵۵۱۲ (۲۷۶/۵۰۰)	۱/۸۸۱۴ (۴۴۲/۵۰۰)	کیسه کلمات + سروازه + برجسب نقش گفتاری سروازه
-	۱/۵۰ (۲۵۰/۵۰۰)	۱/۹۱ (۴۵۵/۵۰۰)	کیسه کلمات + سروازه + برجسب نقش گفتاری سروازه + ویژگی مبتنی بر قانون

همان‌طور که پیش از این نیز اشاره شد، دادگان UIUC تاکنون در پژوهش‌های متعددی مورد استفاده قرار گرفته است. از این رو می‌توان نتایج به‌دست‌آمده از پژوهش جاری را با آن‌ها مقایسه نمود. این نتایج در

قرار می‌دهد. حال آن‌که در پژوهش‌های اخیر، دسته‌بندی ریزدانه‌ها به ازای همه دسته‌های ریزدانه انجام شده است.

جدول ۱۱: صحت دسته‌بندی سایر پژوهش‌های مرتبط که به طور مشابه از دادگان UIUC در ارزیابی‌های خود استفاده کرده‌اند

پژوهش	مدل دسته‌بندی	انواع ویژگی	صحت در دسته‌بندی درشت‌دانه	صحت در دسته‌بندی ریزدانه
Silva et al. [6]	Rule-Based Linear SVM	- نحوی و معنایی	٪۸۷	٪۸۳/۲
Li and Roth [13]	Hierarchical SNoW Hierarchical SNoW	نحوی نحوی / معنایی	٪۹۱	٪۸۴ ٪۸۹/۳
Hacioglu et al. [10]	SVM + ECOC	دوگرام‌ها / n-گرام‌ها	-	٪۸۲ ۸۰-۲۲
Zhang and Lee [11]	Linear SVM SVM tree kernel	n-گرام‌ها n-گرام‌ها	٪۸۷/۴	٪۷۹/۲ ٪۸۰-۲۲
Huang et al. [9]	Linear SVM Maximum Entropy	نحوی / معنایی نحوی / معنایی	٪۹۳/۴ ٪۹۳/۶	٪۸۹/۲ ٪۸۹
Mishra et al. [4]	Linear SVM Naive Bayes K-Nearest Neighbor	نحوی نحوی نحوی	٪۸۸/۲ ٪۷۸/۶ ٪۷۵/۲	-
Hardy and Cheah [7]	ELM	نحوی / معنایی	٪۹۱/۸	٪۸۶/۴
Proposed method	Linear SVM Twin SVM	نحوی + استفاده از مجموعه قوانین ساختاری	٪۸۷/۴ ٪۹۱	٪۷۷/۲ ٪۸۸

پرسش، سرواژه پرسش، برچسب نقش گفتاری پرسش و نوع پاسخ مورد انتظار پرسش باشد. ضمناً، مقایسه میان عملکرد نسخه‌های مختلف SVM نشان داد که الگوریتم Twin SVM از صحت و کارایی بالاتری نسبت به SVM در دسته‌بندی پرسش برخوردار است.

در آینده می‌توان مجموعه قوانین را با در اختیار داشتن دادگان استاندارد بیش‌تر، تکمیل نمود. ضمن این‌که، می‌توان برای مجموعه قوانین، سلسله‌مراتبی در نظر گرفت تا میزان شباهت پرسش‌ها، بر اساس موقعیت آن‌ها در سلسله‌مراتب سنجیده شود. همچنین با توجه به موفق بودن الگوریتم‌های ME و ELM می‌توان رویکردهایی به‌منظور بهره‌مندی از مزایای این الگوریتم‌ها در کنار یکدیگر ارائه داد. علاوه بر این موارد، پیشنهاد می‌شود پیش از ساخت بردارهای کیسه کلمات و به‌منظور کاهش ابعاد، ابتدا ریشه‌یابی از کلمات به عمل آید. ضمن اینکه می‌توان با استفاده از وردنت شباهت میان نام دسته‌ها و سرواژه پرسش را محاسبه نموده و عنوان دسته‌ای را که بیش‌ترین شباهت را به سرواژه دارد، به‌عنوان یک ویژگی جدید به بردار ویژگی افزوده و نقش آن را در صحت دسته‌بند ارزیابی نمود.

۵- نتیجه‌گیری و پژوهش‌های آتی

دسته‌بندی پرسش یکی از مؤلفه‌های حیاتی پاسخ‌گویی به پرسش و تعیین‌کننده کیفیت و کارایی کل سیستم است. از این‌رو، بسیاری از پژوهشگران در تلاش هستند تا پاسخ مناسبی برای چالش‌های موجود در این حوزه بیابند. هدف ما از انجام این پژوهش، غنی‌سازی بردار ویژگی کیسه کلمات حاصل از پرسش‌های برچسب‌دار با استفاده از مجموعه‌ای از قوانین ساختاری است. همچنین با هدف افزایش صحت فرآیند دسته‌بندی، نسخه‌های مختلف الگوریتم SVM را نیز مورد ارزیابی قرار دادیم. نتایج ارزیابی‌ها نشان داد که کلمات توقف موجود در پرسش در تشخیص صحیح دسته معنایی پرسش بسیار مؤثر هستند. علاوه بر این تبدیل صورت نوشتاری تمام کلمات به حروف کوچک، باعث افزایش صحت دسته‌بندی می‌شود. از همه مهم‌تر اینکه رویکرد پیشنهادی این مقاله مبنی بر ترکیب روش‌های مبتنی بر قانون و یادگیری سبب رفع نواقص هریک و بهره بردن از دستاوردهای هر دو است و صحت فرآیند دسته‌بندی به میزان قابل توجهی افزایش می‌یابد. به این ترتیب بهترین بردار ویژگی برداری است که متشکل از کلمات

مراجع

- [1] O. Kolomiyets and M. F. Moens, "A survey on question answering technology from an information retrieval perspective," *Information Science*, vol. 181, no. 24, pp. 5412-5434, 2011.
- [2] T. Desai and A. Parmar, "A survey on different approaches of question answering system," *Advanced Research in Engineering Science and Management*, vol. 1, no. 5, 2013.
- [3] S. K. Dwivedi and V. Singh, "Integrated question classification based on rules and pattern matching," *International Conference on Information and Communication Technology for Competitive Strategies (ICTCS)*, pp. 1-7, 2014.
- [4] M. Mishra, V. K. Mishra and Dr. H. R. Sharma, "Question Classification using semantic, syntactic and lexical features," *International Journal of Web and Semantic Technology*, vol. 4, no. 3, pp. 39-47, 2013.
- [5] B. Loni, *A Survey of State-of-the-art Methods on Question Classification*, Technical Report, Delft University of Technology, pp. 1-40, Aug. 2011.
- [6] J. Silva, L. Coheur, A. C. Mendes and A. Wichert, "From symbolic to sub-symbolic information in question classification," *Artificial Intelligence*, vol. 35, no. 2, pp. 137-154, 2010.
- [7] Hardy and Y. Cheah, "Question classification using extreme learning machine on semantic features," *Journal of ICT Research and Applications*, vol. 7, no. 1, pp. 36-58, 2013.

^{۱۷} WordNet

^{۱۸} Part of Speech (POS)

^{۱۹} Named Entities (NE)

^{۲۰} Singular Value Decomposition (SVD)

^{۲۱} Extreme Learning Machine (ELM)

^{۲۲} Feed Forward (FF)

^{۲۳} Back propagation (BPP)

- [8] U. Hermjakob, E. Hovy, C. Lin and L. Gerber, "Automated question answering in Webclopedia: a demonstration," *2nd International Conference on Human Language Technology Research*, pp. 370-371, 2002.
- [9] Z. Huang, M. Thint and Z. Qin, "Question classification using Headwords and their Hypernyms," *Empirical Methods in Natural Language Processing*, pp. 927-936, 2008.
- [10] K. Hacıoglu and W. Ward, "Question classification with Support Vector Machines and error correcting codes," *Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, vol. 2, pp. 28-30, 2003.
- [11] D. Zhang and W. S. Lee, "Question classification using Support Vector Machines," *The ACM SIGIR Conference in Information Retrieval*, pp. 26-32, 2003.
- [12] V. Krishnan, S. Das and S. Chakrabarti, "Enhanced answer type inference from questions using sequential model," *Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 315-322, 2005.
- [13] A. Sangodiah, M. Muniandy and L. E. Heng, "Question classification using statistical approach: A complete review," *Theoretical and Applied Information Technology*, vol. 71, no. 3, pp. 386-395, 2015.
- [14] P. Biswas, A. Sharan and R. Kumar, "Question classification using syntactic and rule-based approach," *Advanced in Computing Communication and Informatics (ICACCT)*, pp. 1033-1038, 2014.
- [15] C.C.T. Kwok, O. Etzioni, D.S. Weld, "Scaling question answering to the web," *10th International Conference on World Wide Web*, pp. 150-161, 2001.
- [16] N. Kalchbrenner, E. Grefenstette and P. Blunsom, "A Convolutional Neural Network for Modelling Sentences," *52nd Annual Meeting of the Association for Computational Linguistics*, pp. 1-11, 2014.
- [17] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP*, pp. 1746-1751, Sep. 2014.
- [18] http://www.isi.edu/natural-language/projects/webclopedia/Taxonomy/taxonomy_toplevel.html (2015/06/20)
- [19] X. Li, D. Roth, "Learning question classifiers: The role of semantic information," *19th International Conference on Computational Linguistics, Association for computational linguistics*, pp. 1-7, 2002.
- [20] [http://cogcomp.cs.illinois.edu/Data/OA/QC/\(2015/06/20\)](http://cogcomp.cs.illinois.edu/Data/OA/QC/(2015/06/20))

زیرنویس‌ها

^۱ Question Classification (QC)

^۲ Information Retrieval (IR)

^۳ Question Answering (QA)

^۴ Rule-Based

^۵ Machine Learning (ML)

^۶ Accuracy

^۷ Bag-Of-Words (BOW)

^۸ Support Vector Machine (SVM)

^۹ Decision tree (DT)

^{۱۰} Naive Bayes (NB)

^{۱۱} Sparse Network of Winnows (SNOW)

^{۱۲} Headword

^{۱۳} Hypernym